November 20th, 2018

Processing Data Where It Makes Sense in Modern Computing Systems: Enabling In-Memory Computation

Onur Mutlu
ETH Zurich, Switzerland

Abstract
Today's systems are overwhelmingly designed to move data to computation. This design choice goes directly against at least three key trends in systems that cause performance, scalability and energy bottlenecks: 1) data access from memory is already a key bottleneck as applications become more data-intensive and memory bandwidth and energy do not scale well, 2) energy consumption is a key constraint in especially mobile and server systems, 3) data movement is very expensive in terms of bandwidth, energy and latency, much more so than computation. These trends are especially severely-felt in the data- intensive server and energy-constrained mobile systems of today. At the same time, conventional memory technology is facing many scaling challenges in terms of reliability, energy, and performance. As aresult,

memory system architects are open to organizing memory in different ways and making it more intelligent, at the expense of slightly higher cost. The emergence of 3D-stacked memory plus logic as well as the adoption of error correcting codes inside the latest DRAM chips are an evidence of this trend.

In this talk, I will discuss some recent research that aims to practically enable computation close to data. After motivating trends in applications as well as technology, we will discuss at least two promising directions: 1) performing massively-parallel bulk operations in memory by exploiting the analog operational properties of DRAM, with low-cost changes, 2) exploiting the logic layer in 3D-stacked memory technology in various ways to accelerate important data- intensive applications. In both approaches, we will discuss relevant cross-layer research, design, and adoption challenges in devices, architecture, systems, applications, and programming models. Our focus will be the development of in-memory processing designs that can be adopted in real computing platforms and real data-intensive applications, spanning machine learning, graph processing and genome analysis, at low cost. If time permits, we will also discuss and describe simulation and evaluation infrastructures that can enable exciting and forward-looking research in future memory systems, including Ramulator and SoftMC.

Short bio

Onur Mutlu is a Professor of Computer Science at ETH Zurich. He is also a faculty member at Carnegie Mellon University, where he previously held Strecker Early Career Professorship. His current broader research interests are in computer architecture, systems, and bioinformatics. He obtained his PhD and MS in ECE from the University of Texas at Austin and BS degrees in Computer Engineering and Psychology from the University of Michigan, Ann Arbor. His started the Computer Architecture Group at Microsoft Research (2006- 2009), and held various product and research positions at Intel

Corporation, Advanced Micro Devices, VMware, and Google. He received the inaugural IEEE Computer Society Young Computer Architect Award, the inaugural Intel Early Career Faculty Award, US National Science Foundation CAREER Award, Carnegie Mellon University Ladd Research Award, faculty partnership awards from various companies, and a healthy number of best paper or "Top Pick" paper recognitions at various computer systems and architecture venues. He is an ACM Fellow "for contributions to computer architecture research, especially in memory systems" and an elected member of the Academy of Europe (Academia Europaea). For more information, please see his webpage at https://people.inf.ethz.ch/omutlu/.